An Evaluation of Operator Workload, During Partially-Autonomous Vehicle Operations

Salvatore P. Schipani, Ph.D.
U.S. Army Research Laboratory,
Aberdeen Proving Ground, Maryland 21005

ABSTRACT

A series of three field experiments were conducted to evaluate the level of autonomous mobility for the Army's Experimental Unmanned ground Vehicle (XUV), during which an assessment of operator mental workload was performed. Workload data collection methods employed were the "NASA Task Load Index", the "Overall Workload" scale, and experimenter observation during partially autonomous operations, conducted within relevant operating environments including open arid, vegetated, and urban terrains. Although the vehicle was able to successfully traverse terrain unaided at a rate of almost 95%, the level of mental workload increased significantly during periods when human intervention became necessary. Terrain difficulty revealed the most significant effect, followed by an effect for distances traveled. Topography changes, resulting from inclement weather, caused unexpected increases in workload during voyages over what was thought to be the less difficult terrain, articulating the benefit for conducting tests within actual environments in exposing critical operational issues. Additionally, perceived mental workload was highly influenced during conduct of the more deliberate type missions (cautious approches to and from points of interest). The NASA-TLX subscale categories "Temporal" (the amount of time pressure felt), followed closely by "Mental" (the degree of recalling or calculating require), revealed highest workload demand, and comparison of NASA-TLX "Global" ratings with the "Overall Workload" collection method show acutely high correlation, demonstrating the latter an advantageous less obtrusive collection technique. Though the information exposed may be considered most beneficial as baseline performance criterion, it is reasonable to anticipate that as future operators become expected to perform ancillary assignments, reserve human mental capacities should logically decrease.

KEYWORDS: *mental workload measures, robotic, teleoperation, autonomous vehicle.*

1. INTRODUCTION

The United States Army intends to field a Future Combat System (FCS) equipped unit of action (UA) by the end of the decade. One facet in achieving this goal will be a decision milestone when the level of autonomous mobility available to contemporary robotic platforms within the U.S. Amy's Experimental Unmanned Vehicle (XUV) program will be ascertained. Since complete autonomy cannot be achieved at this time and partial autonomy assumes human (operator) intervention, a logical need arose for determining the degree of operator intervention (in the form of workload), required for monitoring or controlling the programs currently most advanced unmanned vehicle.

Natural environments impose significant obstacles to successful navigation by remote systems. Part of the difficulty in attaining complete autonomy lies in the inability of available techniques, especially those involved in sensory interpretation, to classify contextual information and stored knowledge for later recognition of objects and environmental features. Almost two decades ago autonomous robots were designed per a sense-plan-act (SPA) paradigm, where the robot attempted to interpret sensory input with respect to an internal model (based on innate knowledge concerning the environment). Unfortunately, necessarily intense processing costs resulted in limited intelligence, and natural world environment models were often found inadequate [1] [4]. Recently (1999), Brooks [1] proposed a "subsumption architecture" of a set of elementary behaviors activated via external stimuli. In this, stimuli requiring immediate responses (such as a sensor perceived obstacle) evoke appropriate though simple reactions, whereas more complex behaviors (such as exploration) would be performed when simple reactions are inappropriate or else unavailable. This latter behavioral procedure should feasibly be assigned to a human controller. This will also, most likely, result in higher instances of intervention.

The type of human intervention required in assisting partially autonomous platform may be expressed by considering the differences between classical and reactive planning. In a classical, non-reactive mode, a path must be pre-planned as a

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **SEP 2003**	2. REPORT TYPE	3. DATES COVERED **00-00-2003 to 00-00-2003**

4. TITLE AND SUBTITLE **An Evaluation of Operator Workload, During Partially-Autonomous Vehicle Operations**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **U.S. Army Research Laboratory,Aberdeen Proving Ground,MD,21005**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
PerMIS?03, Performance Metrics for Intelligent Systems, 16-18 Sep 2003, Gaithersburg, MD

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**8**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

sequence of coordinate points on a map. This approach requires an extremely accurate model of elements in the world, and the assumption is that all actions produce the desired effect. Reactive planning, on the other hand, requires current sensory information about the remote vehicle, as well as the ability to intervene with new movement control commands. A human may successfully provide three-dimensional connectivity to sensory information in context, essentially in real time. This allows scene feature estimates to be made while in motion, and the location being commanded to can then be designated by the operator based on snap-shot sensor data relayed from the vehicle.

The following is a report of data collected from a series of field experiments comprising a joint agency effort. The primary purpose of the experiment was to evaluate autonomous mobility for one Army candidate partially autonomous vehicle designated the XUV, which utilizes the subsumption architecture mentioned above. An assessment of operator mental workload during mission conduct is the focus of this document. Relevant environments for the entire series of three experiments included: (1) open rolling arid, (2) mixed open rolling vegetated, and (3) urban terrains.

1.1 *Mental Workload Measurement*

Mental workload can be described as the feeling of psychological effort, or the perceived level of use of a human's limited resources. This is typically considered a relative concept, the ratio of demand to available or allocated resources. As task demands increase, resources left in reserve decrease, depending on the input channel involved (*i.e.*, visual or auditory), the degree of processing complexity (cognitive), and the response requirements (psychomotor). Recorded observations of mental workload can take the form of task performance (either primary or secondary), subjective measures, or physiological measures.

Primary task measures may be recorded as task time or accuracy, but concerns for such recordings must be addressed. One potential difficulty in interpreting time or accuracy measures is that the relationship between performance and mental workload is not necessarily linear. Operator self-ratings (subjective techniques) have been shown direct indicators of operator workload [5], and are among the least intrusive. Although best if administered during task performance, these may effectively be administered after the task is complete, therefore resulting in less disturbance during performance. For the current effort, it was not feasible to use a secondary task measure or to record physiological changes, as operations in the given field environments negated utilization of the data collection apparatus necessary. Thus, the measures collected were perceived human workload exertion recordings via subjective techniques, imposed during experimental test runs and subsequently.

The primary objective of the effort detailed within was the quantification of the degree of human workload exerted during the operation of a partially autonomous vehicle, under varying environmental conditions. It was hypothesized that workload would be observed highest during periods of human involvement (rather than mere vehicle supervision), and increase as a function of terrain and mission characteristics.

2. METHOD

2.1 *Participants*

Although a total of eight persons acted as test participants during the entire test sequence, the six soldier-participants of this group only participated during experimental "excursions" (defined later in text). The two remaining participants, civilians, performed for the main effort reported here. They had no prior military experience, however their previous computer usage averaged 5 years of daily use for work and recreation. One civilian technician possessed approximately 1 year of experience teleoperating unmanned ground vehicles (UGVs) as part of a separate program, and the second technician operator held approximately 3 weeks teleoperation experience on the system used for these field tests prior to the beginning of experiments.

2.2 *Apparatus*

<u>National Air and Space Administration – Task Load Index</u>*:* The National Air and Space Administration - Task Load Index (NASA-TLX) [3] is a validated, multidimensional workload rating scale (completed by test participant). This is multidimensional in that it provides information on various sources of workload. The instrument obtains ratings of workload on a scale from low to high for the following six dimensions: (1) mental demand; (2) physical demand; (3) temporal demand; (4) performance; (5) effort, and; (6) frustration. Also produced is a global workload estimate, calculated as the weighted average of the six sub-scale ratings, combining individual ratings into one score. Diagnosticity (referring to the extent to which the specific source or cause of

workload is revealed by the measurement technique) is considered high.

Overall Workload Scale: The Overall Workload (OW) scale [6] produces only an estimate of overall workload, thus is one-dimensional. This is a validated technique, and less obtrusive in administration than the NASA-TLX. The technique obtains a rating of overall workload experienced directly from the operator on a one-dimensional scale from zero to 100 (low to high workload). Once an experimenter establishes a test participant workload profile, the experimenter may assign values as appropriate in lieu of requesting participant responses. Given that a test participant population is small, a profile may be derived by initially observing operations for several hours utilizing OW scoring, and then comparing the resultant workload estimates to that reported from another measure (such as the NASA-TLX).

Experimental Unmanned ground Vehicle (XUV) and Supplementals: The FCS Armed Reconnaissance Vehicle (ARV) representative for this test was the XUV surrogate, a 3500 pound unmanned ground four-wheel steering and drive vehicle powered by a 78-horsepower turbocharged diesel engine, possessing hydrostatic transmission (see Figure 1).

Figure 1. The Experimental Unmanned ground Vehicle (XUV).

As a surrogate, the solitary purpose of this platform's design is to transport sensors. The XUV camera array consisted of one color camera mounted in the sensor pan/tilt pod mechanism on top of the platform, a second black-and-white fixed forward looking camera placed just under the pan/tilt and attached to the vehicle front end, two black and white cameras placed on either side of the vehicle looking forward to allow the operator an indication of platform sides, and a final black and white camera placed low on the rear of the vehicle, culminating in a total of five cameras.

A High-Mobility Multi-purpose Wheeled Vehicle - Control Vehicle (HMMWV-CV) was a second vehicle used for this test (separate from the XUV), which housed the XUV operator (controller), a driver, and the experimenter observer. A third vehicle, the "Safety" HMMWV, traversed in between the XUV and operator HMMWV-CV, shadowing the XUV and prepared to activate a controlled stop of the experimental vehicle if necessary.

Operational Control Unit (OCU): This was the interface device used by XUV operators for monitoring XUV position and system functions via sensor data, and for interacting with the experimental vehicle utilizing computer screen and keyboard input mode (see Figure 2). The OCU screen displayed digital map terrain, vehicle speed, XUV location, and information pertinent to vehicle system's conditions. This also displayed information from three outboard navigational sensors: (1) laser radar; (2) a stereo color camera, and; (3) stereo forward looking infra red.

Figure 2. Windows of the Operator Control Unit (OCU) screen.

Status messages could include a request for help by the XUV when specific (predetermined) conditions occurred, such as when the vehicle autonomously backed up three times or more yet still could not create a good plan for moving beyond an obstacle. During such instances, the operator could take control of (teleoperate) the XUV by activating a secondary video screen containing images sent from one of the cameras mounted on board the XUV, and then utilize the joystick control provided to maneuver the XUV around obstacles. The XUV operator was permitted to intervene only under 1 of approximately 15 tightly defined conditions established for test purposes.

Test Courses: Several test courses were developed at each site to fit the experimental design

and enable multiple simultaneous trials. Independent military subject matter experts devised courses so as to be militarily relevant in terms of tactical movement and objectives. Tooele Army Depot, Utah was selected for rolling arid terrain testing (Phase I). At this site, the course deemed less difficult (designated "Gold") was generally characterized by open terrain and sage brush as high as one meter. The course deemed more difficult (designated "Black") consisted of two distinctly different terrain types, one similar to "Gold" (open terrain) and a second possessing ravines.

Ft. Indiantown Gap Military Reservation, Pennsylvania, was selected for both rolling vegetated wooded and urban terrains (Phases II and III). When tested in rolling vegetated terrain, stretches of tank trails passing through wooded areas interspersed with open terrain were traversed. Characterizing the more difficult area during this test phase was primarily cross-country operation through woods of varying density, open terrain with some dead vegetation mixed with sapling growth, and short stretches of trail containing constriction points such as bridges.

For use during urban testing, one distinct district was established on the Ft. Indiantown Gap Military Reservation. This encompassed a 500-by 200-meter (roughly 2 blocks by 4 blocks) area of one and two-story buildings, possessing through streets, telephone poles, fire hydrants, porches, and concrete barriers (such as those housing coal bins and grease traps), with paths between buildings. For the more difficult test condition at this site, additional temporary barriers were installed, including discarded automobiles, wood and gravel rubble debris piles, and human form mannequins.

2.3 *Procedure*

The XUV performed assigned missions with start and end points preplanned and loaded into the Operator Control Unit. The platform proceeded autonomously until mission completion, or until a determination was made that operator intervention was required through observation of a situation displayed on the OCU monitor (no pre-mission planning was performed). Operator workload was measured at appropriate intervals. During each test run, data were collected during (using the Overall Workload scale) and immediately after (using the NASA-TLX) mission conduct.

2.4 *Experimental Design*

This was a 2 (terrain) *x* 3 (mission) *x* 2 (speed) *x* 2 (offset) repeated measures factorial design, with the factors: (a) *Terrain* (greater difficulty designated "Black" course or "Rubble" when in urban terrain, and less difficult designated "Gold" or "Clean" when in urban terrain); (b) *Mission* (three mission distances traversed being 500, 1000, and 2000 meters, except in the case of urban terrain where distances were necessarily reduced); (c) "offset" designated *Line-of-Sight* (*Line-of-Sight* vs. *Non Line-of-Sight*, as either the Driver of the HMMWV housing the operator being within line of sight of the XUV so that he may offer visual directions to assist during operator intervention, or not), and; (d) *Speed* (an established maximum high speed of 10 meters per second and a low of 3 meters per second, except in the urban terrain where speeds were held at 4 and 2 meters per second).

Each of two teams (consisting of XUV, operator vehicle with personnel, and safety vehicle with personnel) attempted 12 runs per day, totaling 24 runs per test day, which amounted to approximately 10 testing days at each site. Experimental design "excursions" (additional test runs made due to significant scientific interest) included factors for *Light* conditions (day versus night operation), and *soldier-operator* performance comparison with trained technicians, however a description of results is beyond the scope of this report. Normalized (employing a scale of from zero designating low, to 10 designating high) raw workload scores were used in analyses other than when assessing individual NASA-TLX sub-scale ratings. This method was used because of possible confounds associated with the use of previously weighted predictors, and for clarity in comparison across factors.

The derived Overall Workload scale score was the mean of approximately 5 to 10-second workload values reported over the duration of each segment during intervention (though less frequently otherwise), which created a base estimate. If test participants did not respond to experimenter prompts for workload values during instances of teleoperation, this was assumed operator over-load and these periods rated as maximum workload.

3. RESULTS

During Phase I testing, 45 operator interventions were recorded on the Black course while only three were recorded on the Gold, for a total of 48 interventions in all. Here, 171 of 177 missions were successfully completed at 98.3% autonomy. During Phase II testing, 67 operator interventions were recorded on the Black course while 110 were recorded on the Gold, for a total of 177 interventions in all. Here, 155 of 181 missions

were successfully completed at 93.5% autonomy. During Phase III testing, 58 operator interventions were recorded on the Rubble course while 48 were recorded on the Clean, for a total of 106 interventions in all. Here, 264 of 288 missions were successfully completed at 91.7% autonomy.

Of the total 646 main effect experimental test runs made over three terrain environments, although the semi-autonomous vehicle was able to traverse terrain approximately 94.5% of the time on average unaided, when human intervention was necessary operator workload consistently increased (see Figure 3). Across all terrain types, operator interventions were found to significantly affect NASA-TLX "Global" perceived workload ratings ($p < 0.0001$), and rates of intervention correlate highly with increases in workload (0.722).

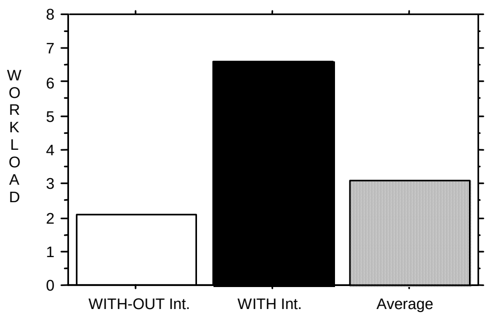

Figure 3. Plot of With-Out and With Intervention, including Average Workload, all test Phases.

During Phase I testing, the mean workload recorded during periods of intervention was 5.8 (of a possible 10), while mean workload during periods without intervention was 1.6 (significantly different, $p < 0.0001$). Total mean workload recorded at this site was 2.5. During Phase II testing, mean workload during periods of intervention was 7.75, while the mean during periods without intervention was 2.4 (significantly different, $p < 0.0001$). Total mean workload recorded at this site was 4.5.

During Phase III testing, mean workload during periods of intervention was 6.1, while mean workload at periods without intervention was 1.9 (significantly different, $p < 0.001$). Total mean workload recorded at this site was 2.8. For combined (all) sites, average workload during periods of intervention was 6.55 (out of a possible 10), while combined average workload during periods without intervention was 1.97 (significantly different, $p < 0.0001$). Combined, the total average workload recorded among all test sites was 3.26.

<u>Phase I, Tooele Army Depot, Utah (rolling arid test site) specific</u>: Neither the *Line-of-Sight* versus *Non Line-of-Sight* independent variable, nor the variable *Speed*, had a statistically significant effect on perceived operator workload (with no interactions found present) ($F[1,72] = 0.014$, $p < 0.9046$ and $F[1,72] = 2.287$, $p < 0.1349$ respectively). The independent variable *Terrain* (difficulty) was found to have significant effect (an increase) on operator's perception of their degree of workload ($F[1, 72] = 5.600$, $p < 0.0207$). "Global" TLX means scores support this finding, and highest averaged perceived workload was seen to take place when traversing the most difficult *Terrain* during the 2000 meter *Mission* (see Figure 4). As for the independent variable *Mission*, though no main effect was revealed for this variable, a significant effect was displayed *post hoc* (Scheffe' test, critical difference = 0.821, $p < 0.0529$) revealing an increase in perceived workload between the two *Mission* distances of 500 and 2000 meters. While remaining effects for the *Mission* variable were slight, the longer distances appeared to increase workload, though to non-significant levels.

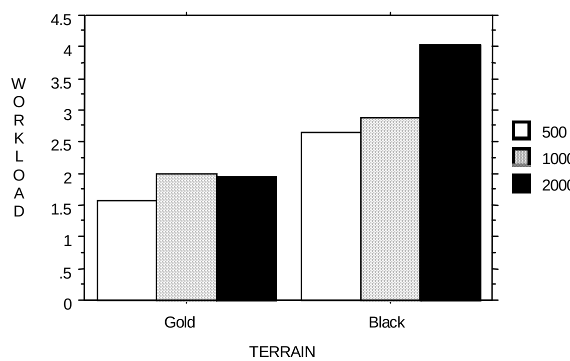

Figure 4. Interaction Cell Plot, *Terrain* by *Mission*, Phase I test.

<u>Phase II, Ft. Indiantown Gap, Pennsylvania (rolling wooded test site) specific</u>: As had occurred during Phase I testing, neither the *Line-of-Sight* versus *Non Line-of-Sight* independent variable, nor the variable *Speed*, had a statistically significant effect on perceived operator workload ($F[1, 140] = 0.204$, $p < 0.652$ and $F[1, 140] = 0.163$, $p < 0.121$ respectively). Again, the independent variable *Terrain* (difficulty) was found to have significant effect on operators' perception of their degree of workload ($F[1, 138] = 3.948$, $p < 0.0489$). However, highest averaged perceived workload was seen to take place when traversing the <u>least</u> difficult *Terrain* during this phase of testing,

especially during the 2000 meter *Mission*, which divulged terrain anomaly (see Figure 5).

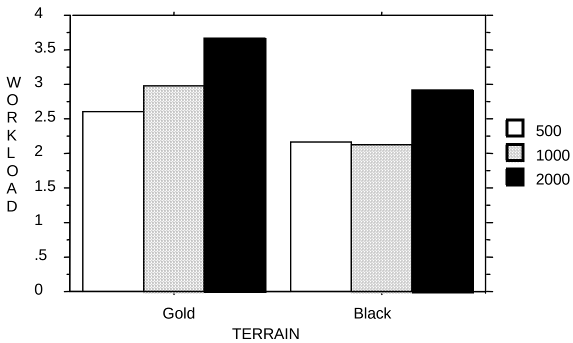

Figure 5. Interaction Cell Plot, *Terrain* by *Mission*, Phase II test.

As for the independent variable *Mission*, a significant effect was displayed as an increase in perceived workload, once again greatest between the two *Mission* distances of 500 and 2000 meters. This was supported in *post hoc* testing ($p < 0.0337$). Though remaining effects for the *Mission* variable were slight, the longer distances continued a trend of increasing workload regardless of terrain anomalies.

<u>Phase III, Ft. Indiantown Gap, Pennsylvania (urban test site) specific:</u> Because criteria other than mere distance were deemed of greater importance for urban operations, modifications were made in the experimental design during this final test sequence. The factor *Mission*, previously with levels of 500, 1000, and 2000 meters, was made more substantive to urban operations by assigning levels of "Patrol" and "Attack" type missions. A Patrol mission is one in which the vehicle must follow designated pathways, acting cautiously in deliberately approaching each way-point. An Attack posture is one in which the vehicle would seek the straightest path between way-points, proceeding directly and immediately. Levels of the factor *Speed* were reduced from maximums of 3 and 10 meters per second to 2 and 4 respectively, as subject matter experts agreed these were reasonably achievable and support successful urban mission performance.

As had occurred consistently during test Phases I and II, neither the *Line-of-Sight* versus *Non Line-of-Sight* independent variable, nor the variable *Speed*, had a statistically significant effect on perceived operator workload ($F[1, 224] = 2.16$, $p < 0.145$ and $F[1, 224] = 0.014$, $p < 0.906$ respectively). Once again, the independent variable *Terrain* was found to have significant effect on operator's perception of workload ($F[1, 224] = 6.794$, $p < 0.0107$), as higher averaged perceived workload was seen to take place when traversing the more difficult *Terrain* as had been predicted (see Figure 6). Concerning the independent variable *Mission*, a significant effect was displayed as an increase in perceived workload ($F[1, 224] = 8.163$, $p < 0.0053$), however seen highest during the "Patrol" type mission. Apparently, the more deliberate vehicle operation caused by a patrol mission was enough to significantly increase test participants' perception of exertion.

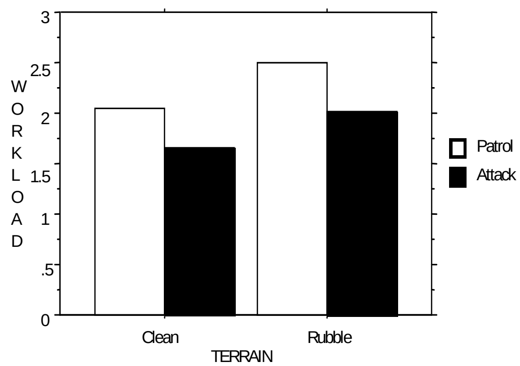

Figure 6. Interaction Cell Plot, *Terrain* by *Mission*, Phase III test.

<u>NASA-TLX Sub-categories, Collapsed over Phases</u>: Reviewing TLX "sub-category" ratings resulting from and collapsed over the current effort, the factor "Temporal", followed closely by sub-category "Mental", showed highest perceived workload demand throughout testing. "Temporal" workload demand reveals the amount of time pressure operators feel when performing duties and accounted for approximately 32.7% of the averaged total workload recorded, while "Mental" demand relates to the amount of thinking, calculating, or remembering required for performance and accounted for approximately 28.6% total workload. The TLX sub-category "Frustration" accounted for approximately 18.7% of averaged workload recorded (this significantly increased during Phase II testing only, where $p < 0.0452$), followed by "Effort" accounting for approximately 12.5%. The subcategory "Physical" consistently accounted for the least amount of overall averaged workload (approximately 5%). Finally, the TLX 'sub-category' "Performance" was consistently found extremely high, displaying feelings of satisfaction with performance.

A statistically significant assessment supported the fact that combined NASA-TLX sub-scales revealed a similar picture as that produced by the

TLX "Global" rating ($p < 0.001$), and comparison of NASA "Global" ratings with data collected via the "Overall Workload" method show acutely high correlation (estimated raw workload averages of 2.46795 and 2.40551 reported respectively, of a possible 10.0 maximum, assessed during Phase II testing). Although data for experimental main effects were collected from only two test participants resulting in achieved low by-participant statistical power (0.053), differences in workload recorded among individual participants (assessed during Phase I, the site with least main effect experimental runs) were not significant ($F[1, 76] = 0.03$, $p < 0.864$).

4. DISCUSSION

Data were collected during a series of field experiments comprising a joint agency effort, the primary purpose of which was to evaluate autonomous mobility as a function of sensor performance for one Army candidate partially autonomous vehicle designated the XUV. A subset of this experiment was the assessment of operator mental workload during mission conduct, which is the focus of this report. Relevant environments for the total of three experiments included open rolling arid, mixed open rolling vegetated, and urban terrains where primarily dismounted actions occur. To determine degrees of workload, three methods of data collection were employed: a) the "NASA Task Load Index" (NASA-TLX) workload scale ratings; b) the "Overall Workload" scale [6], and; c) experimenter observation.

The XUV may be controlled, during periods deemed essential, via teleoperation by a geographically separated user, although maximum autonomy is desirable. This intermediate level of automation (LOA), considered "management by consent" (automation proposes actions but cannot proceed without explicit operator consent), is appropriate because such methodology should support consistent performance as future system complexities increase, and may allow for several vehicles to be controlled by a single supervisor if desired. Throughout the current test series, although the vehicle was able to successfully traverse terrain approximately 94.5% of the time on average unaided, workload increased significantly ($p < 0.0001$) when intervention by the human was necessary, and the rate of intervention correlated highly with workload perceived. For combined (all) sites, averaged workload during periods of intervention was 6.55 (of a possible 10), while the average recorded when no interventions were necessary was at the 1.97 level.

Experimental variables manipulated were: (a) *Terrain* (greater versus less difficult); (b) *Mission* (distances and types traversed); (c) control vehicle to XUV "offset", designated *Line-of-Sight* (*Line-of-Sight* versus *Non-*, as within sight of the XUV so that visualization during intervention could be employed to assist operator, or not), and; (d) *Speed* (established high and low limits). A total of 646 experimental runs were completed between three test sites. Neither the *Line-of-Sight* versus *Non*, nor the independent variable *Speed*, revealed statistically significant effect on operator workload. This may be attributable to the specifics of having too few instances of teleoperation to reveal effects for the former, and inconsequential speed disparity in the latter. The variables *Terrain* and *Mission* did, however, return significant effects throughout testing, in the form of perceived increases in operator workload.

In rolling arid testing (Phase I, Tooele Army Depot, Utah), differences in *Terrain* difficulty revealed a significant effect ($p < 0.0207$), in that the more difficult area at this site increased workload. Similarly, when *Mission* distances traversed were greatest, *post hoc* testing revealed a significant effect ($p < 0.0529$) between shortest and furthest distances traversed.

During rolling wooded testing (Phase II, Ft. Indiantown Gap, Pennsylvania), a significant effect was displayed *post hoc* for the independent variable *Mission* as an increase in perceived workload ($p <$ 0. 0529). Once again, the longer mission distance increased the perception of workload. The variable *Terrain* again produced a significant effect on operators' perception of their degree of workload ($p < 0.0489$), however suprisingly took place during travel over the *less* difficult terrain due to anomolies previously undistinguishable (topography transformations because of rain saturatation, and inclement weather including unexpected snow which caused vehicle slippage). This revelation articulates the benefits of conducting tests in actual operating environments.

Field evaluation generally describes an attempt to gain concept knowledge while performing activities normally involved with utility testing in the context intended, rather than merely confirming a belief. Here, one must tolerate unpredictable conditions, and work within constraints of the environment to gain knowledge about potential integration problems caused by the impact of real issues. Had the current system not been exposed to this more ecological test approach, situations that might not transfer well from the laboratory to a field setting may not have been revealed, thus the

critical operational issues exposed would remain undefined.

At the urban testing site (Phase III, Ft. Indiantown Gap, Pennsylvania), the independent variable *Mission* returned a significant effect ($p < 0.0053$) on the operator's perception of workload, seen as a workload increase during the "Patrol" type mission. Though unexpected, apparently this more deliberate type operation (cautious approaches to and from points of interest, normally assumed during urban operations) increases the level of human effort necessary to perform successfully. As was true during all previous test terrains, the independent variable *Terrain* returned a significant effect on the operator's perception of workload ($p < 0.0107$). As predicted, greater operational difficuly ensued when when the XUV traversed the more debris cluttered terrain, resulting in an increase in the perception of workload.

Aside from producing a "Global" workload estimate, the NASA-TLX workload rating scale provides six dimensions (sub-scale categories) for determining the psycho-physiological loci of emanating workload. Of these, the factor "Temporal" (the amount of time pressure felt when operating), followed closely by the sub-scale "Mental" (the amount of thinking, recalling, or calculating require), revealed highest perceived workload demand throughout testing (averages of 32.7 and 28.6% of the total workload respectively). The TLX sub-scale "Frustration" accounted for approximately 18.7% of averaged workload, significantly ($p < 0.0452$) increasing during Phase II testing only as a function of unexpected (to operators) terrain transformation (during this test Phase, the sub-scale "Effort" also appeared to increase slightly, though never reached a level of significance). The sub-scale "Effort" accounted for approximately 12.5% of the averaged workload perceived, and "Physical" consistently returned the least amount recorded (approximately 5%). "Performance" was consistently extremely high, displaying operators' feelings of satisfaction with accomplishments.

A comparison of NASA-TLX "Global" ratings with data collected via the "Overall Workload" method (estimated raw averages of 2.46795 and 2.40551 reported, respectively) shows acutely high correlation (0.819), demonstrating this to be an advantageous and less obtrusive alternate data collection method possessing the possibility for establishing a performance profile.

The information reported within may be considered for use as baseline performance criterion for operators of partially autonomous vehicles, those who might supervise autonomous operations for the most part though be required to intervene during brief and sporatic periods. This may be especially true for the platform employed during the current experiments, as this performed autonomously at a rate of almost 95%, and test participants were not required to conduct additional duties. However, it is reasonable to expect that as robotic operators are given ancillary assignments such as monitoring systems bearing mission packages, asked to supervise the conduct of multiple robotic platforms, or else required to attempt both, reserve human capacities should logically decrease while rates of workload exerted necessarily increase. One must also consider that without human intervention, any period of vehicle incapacitation most likely equates to mission failure. Thus, to give acute consideration for the human element when contemplating fielding such systems seems imperative.

5. REFERENCES

[1] Braitenberg, V. (1984). *Vehicles: Experiments in synthetic psychology*. Cambridge, MA: MIT Press.

[2] Brooks, R. (1999). *Cambrian Intelligence*. Cambridge, MA: MIT Press.

[3] Hart, S.G. & Staveland, L.E. (1987). Development of a NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P.S. Hancock & N. Meshakti (Eds.), *Human mental workload*. Amsterdam: Elsevier.

[4] Lieblich, I. & Arbib, M. A. (1982). Multiple representation of space underlying behavior. *The Behavioral and Brain Sciences*, *5*, pp. 627-659.

[5] Sheridan, T.B. (1992). *Telerobotics, automation, and human supervisory control*. Cambridge, MA: The MIT Presss.

[6] Vidulich, M.A., & Tsang, P.S. (1987). Absolute magnitude estimation and relative judgment approaches to subjective workload assessment. *Proceedings of the Human Factors Society 31st Annual Meeting* (pp. 1057-1061). Santa Monica, CA: Human Factors Society.